基于改进 CHI 和带权 ECE 结合的特征选择方法 *

蔡　镇，高　健，秦晓军

(江南计算技术研究所，江苏 无锡 214083)

摘　要：针对文本分类特征选择方法中的卡方统计(CHI)和期望交叉熵(ECE)，分析了其特点和不足。为了避免传统 CHI 和 ECE 方法在不平衡数据集上分类效果差的问题，本文通过引入调节因子和除去负相关影响因素，给出了改进的 CHI 方法(pCHI)，并以加权的方式弥补 ECE 方法倾向于选择弱区分能力高频特征的缺陷(ωECE)。在综合两种改进后方法的基础上，进一步提出基于改进 CHI 和带权 ECE 结合(pCHIωECE)的特征选择方法。经对比实验验证，pCHIωECE 方法的查准率、F1 值均优于 CHI、ECE 及 pCHI、ωECE 方法，且该方法的降维稳定性更好。

关键词：卡方统计；期望交叉熵；特征选择；文本分类

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.03.0240

Feature selection method based on combining improved CHI and weighted ECE

Cai Zhen, Gao Jian, Qin Xiaojun

(*JiangNan Institute of Computing Technology*, *Wuxi Jiangsu* 214083, *China*)

Abstract: This paper analyzed the characteristics and deficiencies of chi-square statistics and expected cross-entropy methods for feature selection of text classification. In order to avoid the poor classification of traditional CHI and ECE methods on unbalanced data sets, this paper presented an improved CHI method (pCHI) by introducing adjustment factors and removing negative correlation influencing factors, and presented a weighted ECE method(ωECE) to compensate the disadvantages of the ECE method tending to select high-frequency features of weak distinguishing ability. After synthesizing the two improved methods, this paper further proposed a feature selection method based on combining improved CHI and weighted ECE (pCHIωECE) . Through comparative experiments, the precision rate and F1 value of the pCHIωECE method are both superior to those of the CHI, ECE, pCHI and ωECE methods, and moreover, the dimensionality and stability of the method are better.

Key words: Chi-square statistics; expected cross-entropy; feature selection; text classification

0　引言

文本分类[1]是指将大量的文本按照预先定义的分类体系归到一个或者多个类别的技术，该技术被广泛应用于数据挖掘、机器学习、信息检索等领域。文本分类大致可分为文档表示、特征选择和分类器训练等。特征选择是指从一个原始的特征空间选择一个最优特征子空间的过程。由于文本特征的“维数灾难[2]”以及不相关特征（噪声）的存在，特征选择对于文本分类尤其重要。文本分类常用的特征选择算法是基于信息论和统计学思想设计的，包括基尼指数、文档频率、信息增益、互信息、卡方统计、期望交叉熵、线性判别分析等。文献[3~6]对特征选择的常用方法及特点作了详细阐述。针对传统卡方统计和期望交叉熵方法对不平衡数据集和噪声干扰导致分类效果差的问题，本文提出了一种基于改进 CHI 和带权 ECE 结合的特征选择方法，通过对比实验分析该方法能有效提高文本分类的精度。

1　相关工作

文本分类常用向量空间(VSM)模型表示，设文本表示为 $D=\{X_1,...,X_N\}$，N 是文本总数。$p_1(t),...,p_k(t)$ 表示特征 t 在 k 个不同类别的概率，k 是类别总数，即 $p_i(t)$ 表示文档属于第 i 类中包含特征 t 的条件概率且 $\sum_{i=1}^{k} p_i(t)=1$。P_i 为类别 i 的全局概率，$F(t)$是包含特征 t 的文档数量。

1.1　卡方统计（CHI）

卡方统计[7]是用于度量特征 t 与特定类别 i 之间是否具有非独立性的方式。特征 t 与类别 i 之间的卡方统计量定义为:

$$\chi_i^2(t)=\frac{N\cdot F^2(t)\cdot(p_i(t)-P_i)^2}{F(t)\cdot(1-F(t))\cdot P_i\cdot(1-P_i)} \tag{1}$$

特征t对于全局卡方统计量可以用加权平均值或最大值计算，公式为

收稿日期：2018-03-22；**修回日期：**2018-06-20　**基金项目：**国家自然科学基金资助项目（61732018）

作者简介：蔡镇（1990-），男江苏大丰人，工程师，硕士研究生，主要研究方向为网络安全、自然语言处理（caizhen@mail.ustc.edu.cn）；高健（1981-），女，助理研究员，硕士，主要研究方向为高性能计算；秦晓军（1975-），男，高级工程师，博士，主要研究方向为软件安全.

$$\chi^2_{avg}(t)=\sum_{i=1}^{k}P_i\cdot\chi^2_i(t) \quad (2)$$

$$\chi^2_{\max}(t)=\max_{1\le i\le k}\chi^2_i(t) \quad (3)$$

再按照倒序排列，从而完成特征选择。卡方统计量是一个标准化的值，其值在同类别的各个特征之间具有很强的区分度。若特征t与类别i相互独立，则 $\chi^2_i(t)=0$。特征t与类别i的相关性越强，$\chi^2_i(t)$ 的值就越大，此时特征t所包含的与类别i相关的信息就越多。

1.2　信息增益(IG)与期望交叉熵(ECE)

信息增益[1]通过统计某一个特征t在类别i中是否出现的文档频数来计算特征t对类别i的信息增益，它考虑了特征t出现前后的信息熵之差。特征t的信息增益公式为

$$\begin{aligned}IG(t)=&-\sum_{i=1}^{k}P_i\cdot\log(P_i)\\&+F(t)\cdot\sum_{i=1}^{k}p_i(t)\cdot\log(p_i(t))\\&+F(\bar{t})\cdot\sum_{i=1}^{k}(p_i(\bar{t}))\cdot\log(p_i(\bar{t}))\end{aligned} \quad (4)$$

其中：$F(\bar{t})=1-F(t)$，$p_i(\bar{t})=1-p_i(t)$。信息增益IG(t)的值越大，特征t的区分能力就越大。

期望交叉熵[8]与信息增益相似，不同之处在于ECE只计算出现在文本中的特征，而不考虑特征未出现的情况。给定特征t的期望交叉熵为

$$\begin{aligned}ECE(t)=&-\sum_{i=1}^{k}P_i\cdot\log(P_i)\\&+F(t)\cdot\sum_{i=1}^{k}p_i(t)\cdot\log(p_i(t))\end{aligned} \quad (5)$$

同样，期望交叉熵ECE(t)的值越大，特征t的区分能力就越大。

文献[9,10]表明，IG的特征t不出现也可能对类别判定有贡献，往往是该特征的贡献远小于它所带来的干扰。特别是，在类别和特征分布是高度不平衡的情况下，若 $1-F(t)\gg F(t)$，即绝大多数特征不出现，式(1)中IG(t)的值由 $(1-F(t))\cdot\sum_{i=1}^{k}(1-p_i(t))\cdot\log(1-p_i(t))$ 决定，此时IG(t)更倾向于选择频度小的词。ECE不考虑特征不出现在同类文档中对类别的影响，这正是ECE表现优于IG的原因。

2　基于 pCHIωECE 的特征选择方法

2.1　CHI 分析及改进

由式(1)，CHI方法由 $P_i\cdot F(t)$ 给出不同时满足文本属于类别i且包含特征t的情况，$F(t)\cdot p_i(t)$ 给出同时满足或同时不满足文本属于类别i且包含特征t的情况。事实上，$F(t)\cdot p_i(t)$ 的值比 $P_i\cdot F(t)$ 大或者小取决于类别i和特征t之间的相关程度。

特征与类别包含正相关和负相关这两种情况。令 $\tau=F(t)\cdot p_i(t)-P_i\cdot F(t)$，若 $\tau>0$ 时，有 $p_i(t)>P_i$，特征t与类别i正相关，$\chi^2_i(t)$ 的值越大，文档中包含的特征t属于i类别的可能性越大；反之，若 $\tau<0$ 时，$p_i(t)<P_i$，特征t与类别i负相关，$\chi^2_i(t)$ 的值越大，文档中包含的特征t不属于i类别的可能性越大。传统CHI统计方法只考虑了特征词在所有文档集中出现的文档的数量，而没有考虑特征词在某一篇文档中出现的次数，从而夸大了低频词的作用。在不平衡的样本中，分类效果下降明显。

针对CHI方法夸大低频词的缺陷，本文通过引入特征频率因子α减少低频特征对文本分类的干扰。α的计算公式为

$$\alpha_i(t)=\frac{tf_i(t)}{\sum_{j=1}^{k}tf_j(t)-tf_i(t)+1} \quad (6)$$

其中：$tf_i(t)$ 表示特征t在类别i中出现的频数，$\sum_{j=1}^{k}tf_j(t)$ 表示特征t在所有类别中出现的频数，因子 $\alpha_i(t)$ 表示特征t在某个特定类别i中出现的频数与所有其他类别中出现的频数之比，式（6）分母加1是为了防止发生特征t仅分布在类别i中，以致的分母为0的情况出现。若 $\alpha_i(t)$ 越大，特征t在类别i中出现频数越多，在其他类别中出现频数越少，由此判别特征对类别能够提供更大区分能力。反之，$\alpha_i(t)$ 越小，该特征t的区分能力越小。

再除去其特征与类别的负相关情况，结合公式(1)、（6）改进后的卡方统计公式为

$$\chi'^2_i(t)=\begin{cases}\dfrac{N\cdot F^2(t)\cdot(p_i(t)-P_i)^2}{F(t)\cdot(1-F(t))\cdot P_i\cdot(1-P_i)}\cdot\alpha_i(t), & p_i(t)>P_i\\ 0, & p_i(t)\le P_i\end{cases} \quad (7)$$

2.2　ECE 分析及加权

ECE方法既考虑了特征和类别的相关性，同时也兼顾了特征频率和类别频率之间的差值，但该方法也存在明显不足。由式(5)，如果出现的 $p_i(t)$ 大且 P_i 小的情况，此时特征t对分类的影响大，相应的ECE(t)值就大。由此说明ECE方法没有考虑特征在数据集类间的分布情况，会造成该算法倾向于选择区分能力不强的高频特征[10]。

针对这些不足，本文综合特征出现与否以及类间出现比重

这两个因素，用词频 $tf_i(t)$ 作为权重的计算方式，评价一个特征所含有的信息量。归一化处理后权重计算公式为

$$\omega_i(t)=\frac{tf_i(t)}{\sqrt{\sum_{j=1}^{k}tf_j^2(t)}} \tag{8}$$

结合式(5)（8），加权后的期望交叉熵公式为

$$\begin{aligned}\omega ECE(t)=&-\sum_{i=1}^{k}P_i\cdot\log(P_i)\\&+F(t)\cdot\sum_{i=1}^{k}(p_i(t)\cdot\omega_i(t))\cdot\log(p_i(t)\cdot\omega_i(t))\end{aligned} \tag{9}$$

它反映了文本类别i的概率分布 P_i 与文档属于第i类中包含特征t的条件概率分布 $p_i(t)$ 之间的距离。特征的期望交叉熵越大，对文本分类的影响也就越大。

2.3 pCHIωECE 特征选择方法

通过分析CHI和ECE的特点与不足，本文分别对这两种特征选择方法做了优化处理。整合两种优化后的方法，从而得到基于pCHI和ωECE结合(pCHIωECE)的特征选择方法。

pCHIωECE的设计流程，如图1所示。

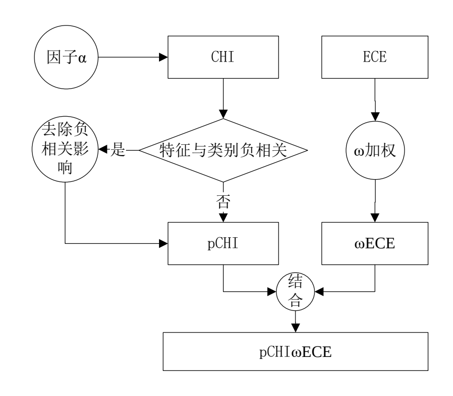


图1 pCHIωECE的设计流程

改进CHI(pCHI)方法，降低了低频特征对文本分类的干扰和去除负相关因素影响，使得pCHI的相较于CHI有更好的降维能力。ωECE方法减轻了ECE方法对区分能力弱的高频特征的倚重，一定程度上提高了ECE的分类效果。pCHIωECE方法结合CHI和ECE各自方法的特点，不仅缓解低频词缺陷，也可选择出集中出现在特定类别且频数高的特征，进而使得分类效果和稳定性更好。pCHIωECE计算公式为

$$\begin{aligned}pCHI\omega ECE(t)&=\chi'^2_{\max}(t)\cdot\omega ECE(t)\\&=\max_{1\le i\le k}\chi'^2_i(t)\cdot\omega ECE(t)\end{aligned} \tag{10}$$

3 实验结果及分析

3.1 实验环境和数据

实验环境为在Windows 10 x64操作系统，Inter(R) Core™ i5-5250U @1.6 GHz处理器，4 GB内存的PC，开发工具为Python3.6。通过调用Python的Sklearn模块编程实现本文引用、改进和提出的5种特征选择方法，选用朴素贝叶斯(NB)分类器完成分类。

数据集来源是复旦大学中文语料库共包含20个类别9833篇，各类文本数分别为空间642，能源33，电子28，通信27，计算机1358，矿产34，交通59，艺术742，环境1218，农业1022，经济1601，法律52，医疗53，军事76，政治1026，体育1254，文学34，教育61，哲学45，历史468。从各类文本分布看，数据集是极不平衡的。本实验将其中的80%作为训练集，20%作为测试集。

3.2 评价指标

本文使用的评价标准是文本分类中普遍使用的性能评价指标[11]：查准率(precision)、查全率(recall)和F1值。对于文本分类问题，根据真实类别与预测类别的组合划分为真正例(TP)、假正例(FP)、真反例(TN)、假反例(FN)四种情形，分类结果的“混淆矩阵”如表1所示。

表1　分类结果的“混淆矩阵”

真实情况	预测结果：正例（预测属于某类别的文本数）	预测结果：反例（预测不属于某类的文本数）
正例(真实属于某类的文本数)	TP	FN
反例(真实不属于某类的文本数)	FP	TN

查准率precision定义为

$$precision=\frac{TP}{TP+FP}$$

查全率recall定义为

$$recall=\frac{TP}{TP+FN}$$

F1是基于查准率和查全率的调和平均定义的，F1的度量为

$$F1=\frac{2\cdot precision\cdot recall}{precision+recall}$$

为了综合考虑查准率和查全率,本文使用F1值作为评价指标。

3.3 实验结果与分析

实验通过将CHI、ECE与pCHI、ωECE以及pCHIωECE进行比较分析，从而验证本文提出的算法分类正确性和性能。

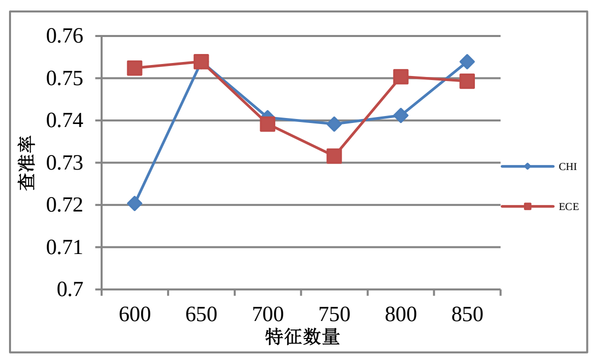


图2　CHI和ECE方法的查准率对比

图2是传统CHI和ECE特征选择的查准率对比，由图2可得，这两种方法在不同特征数量下的查准率有明显波动，ECE整体查准率相对较高，特征数量大于650时，CHI的查准率更稳定。

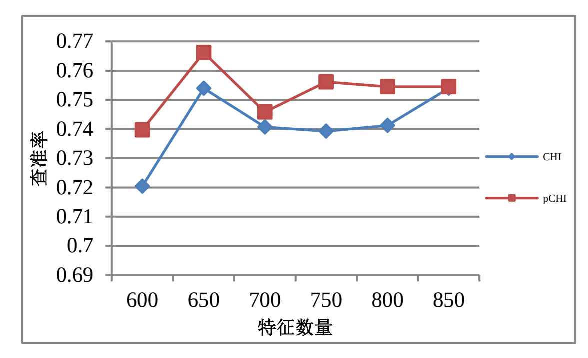


图3　pCHI和CHI方法的查准率对比

图3是改进CHI特征选择方法（pCHI）和CHI方法的准确度对比，pCHI引入了调节因子α和除去特征与类别负相关的情况。由图3可得，pCHI查准率变化趋势与CHI方法有一定的相关性，pCHI的查准率均高于传统的CHI方法，验证了pCHI方法的正确性和优势。

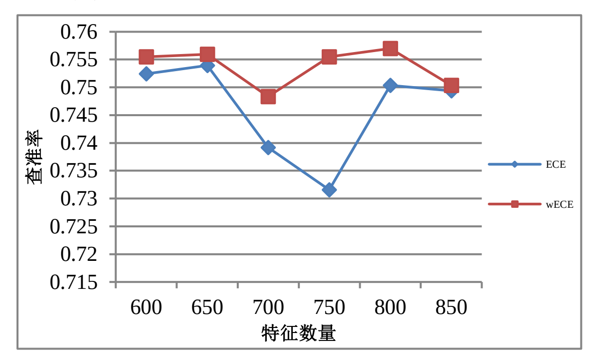


图4　ωECE和ECE方法的查准率对比

图4是带权ECE特征选择方法（ωECE）和ECE方法的准确度对比，ωECE考虑了分类特征出现和类间出现比重这两个因素加权。如图4所示，查准率变化趋势比原有的ECE方法更加稳定，同时ωECE的查准率高于ECE方法，验证了ωECE方法的正确性和优势。

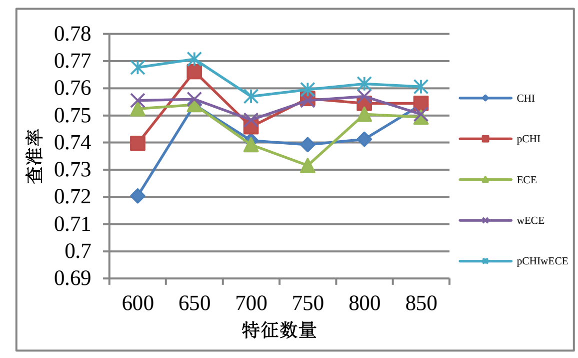


图5　pCHIωECE与其他方法的查准率对比

图5是pCHIωECE与CHI、pCHI、ECE、ωECE的查准率对比曲线。如图5所示，pCHIωECE方法的查准率比其他四个方法都要高。传统的CHI、ECE方法准确度相对较低，但随着特征数量增加有所提升。pCHI、ωECE方法比改进前的查准率略有提升，pCHI与ωECE在不同特征数量上的查准率各有优劣，降维能力相近。从分类结果可以看出,文本数据集类间分布不平衡对分类的结果有一定影响。在特征数量低于700时准确率更高，说明pCHIωECE方法降维能力更好。该方法在整体稳定性较高，一定程度上提升了数据集不平衡条件下的分类效率，整体查准率优于其他方法。

采用贝叶斯分类的五种特征选择方法F1值，如图6所示。

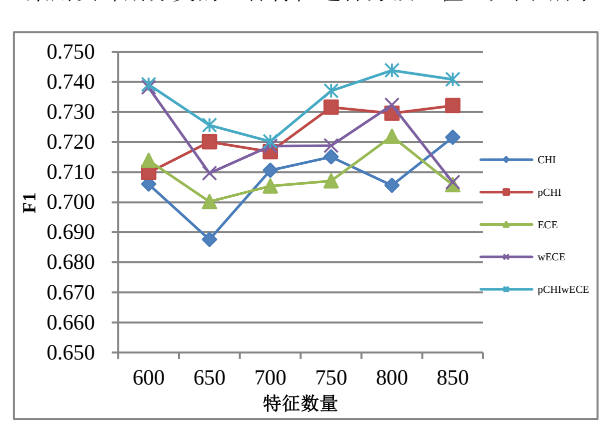


图6　pCHIωECE方法与其他方法的F1值对比

由图6对比分析，CHI与ECE方法的特征数量没有明显相关性，pCHI、ωECE方法的F1值分别高于CHI、ECE的F1值，pCHIωECE方法的F1值比其他方法都要高。

从查准率和F1值的结果看，本文提出的基于改进CHI和带权ECE结合的特征选择方法分类效果优于CHI和ECE方法。

4　结束语

本文首先分析了CHI和ECE特征选择方法的特点和不足，通过引入调节因子和除去负相关影响改进CHI及使用加权的方式优化ECE，在此基础上提出了基于改进CHI和带权ECE结合的特征选择方法。进而在不平衡的数据集上做对比实验，分析了不同方法查准率、F1值，定性表明pCHI、ωECE、pCHIωECE分类方法均有不同程度的提高分类效果，从而验证了pCHIωECE方法能提高文本分类的正确性且稳定性更好。

参考文献：

[1] Aggarwal C C, Zhai ChengXiang. A survey of text classification algorithms [M]// Mining Text Data. Boston: Springer, 2012: 163–222.

[2] Cai Jie, Luo Jiawei, Liang Cheng, *et al.* A novel information theory-based ensemble feature selection framework for highdimensional microarray data [J]. International Journal of Performability Engineering, 2017, 13 (5): 742-753.

[3] Vijayan V K, Bindu K R, Parameswaran L. A comprehensive study of text classification algorithms [C]// Proc of International Conference on Advances in Computing, Communications and Informatics. 2017: 1109-1113.

[4] 毛勇，周晓波，夏铮，等. 特征选择算法研究综述 [J]. 模式识别与人工智能, 2007, 20 (2): 211-218. (Mao Yong, Zhou Xiaobo, Xia Zheng, *et al.* A survey for study of feature selection algorithms [J]. Pattern Recognition and Artificial Intelligence, 2007, 20 (2): 211-218.)

[5] 张群，王红军，王伦文. 基于词条属性聚类的文本特征选择算法 [J]. 计算机应用研究, 2017, 34 (2): 369-372, 377. (Zhang Qun, Wang Hongjun, Wang Lunwen. Algorithm of text freature selection based on vocabulary attribute clustering [J]. Application Research of Computers, 2017, 34 (2): 369-372, 377.)

[6] Rehman A, Javed K, Babri H A. Feature selection based on a normalized difference measure for text classification [J]. Information Processing & Management, 2017, 53 (2): 473-489.

[7] 高宝林，周治国，杨文维，等. 基于类别和改进的 CHI 相结合的特征选择方法 [J]. 计算机应用研究, 2018, 35 (6): 1660-1662. (Gao Baolin, Zhou Zhiguo, Yang Wenwei, *et al.* Feature selection method based on combination of category and improved CHI [J]. Application Research of Computers, 2018, 35 (6): 1660-1662.)

[8] 王海鹏，韩立新，甄志龙. 基于索引项权重的文本特征选择方法 [J]. 计算机工程与设计, 2010, 31 (5): 1149-1151. (Wang Haijuan, Han Lixin, Zhen Zhilong. Feature selection based on term weight for text categorization. Computer Engineering and Design, 2010, 31 (5): 1149-1151.)

[9] 单丽莉，刘秉权，孙承杰. 文本分类中特征选择方法的比较与改进 [J]. 哈尔滨工业大学学报, 2011, 43 (S1): 319-324. (Shan Lili, Liu Bingquan, Sun Chengjie. Comparison and improvement of feature selection method for text categorization [J]. Journal of Harbin Institute of Technology, 2011, 43 (S1): 319-324.)

[10] 杜同森. 文本分类中特征选择和特征加权算法的研究 [D]. 北京：北京邮电大学, 2014. (Du Tongsen. Research on feature selection and feature weighting algorithm in text categorization [D]. Beijing: Beijing University of Posts and Telecommunications, 2014.)

[11] 王振，邱晓晖. 混合 CHI 和 MI 的改进文本特征选择方法 [J]. 计算机技术与发展, 2018, 28 (4): 87-90, 94. (Wang Zhen, Qiu Xiaohui. An improved text feature selection method mixed CHI and MI [J]. Computer Technology and Development, 2018, 28 (4): 87-90, 94.)

[12] Liu Jinghua, Lin Yaojin, Lin Menglei, *et al.* Feature selection based on quality of information. Neurocomputing, 2017, 225: 11-22.

[13] 孟佳娜，林鸿飞，李彦鹏. 基于特征贡献度的特征选择方法在文本分类中应用 [J]. 大连理工大学学报, 2011, 51 (4): 611-615. (Meng Jiana, Lin Hongfei, Li Yanpeng. Application fo feature selection method to text categorization based on feature contribution degree [J]. Journal of Dalian University of Technology, 2011, 51 (4): 611-615.)

[14] 邱云飞，王威，刘大有，等. 基于方差的 CHI 特征选择方法{J]. 计算机应用研究, 2012, 29 (4): 1304-1306. (Qiu Yunfei, Wang wei, Liu Dayou, *et al.* CHI feature selection method based on variance [J]. Application Research of Computers, 2012, 29 (4): 1304-1306.)